

Multiple Alignment of Structures using Center Of ProTeins

Kaushik Roy and Satish Ch. Panigrahi and Asish Mukhopadhyay*

School of Computer Science
University of Windsor
401 Sunset Avenue
Windsor, ON N9B 2R4, Canada
roy113@uwindsor.ca
panigra@uwindsor.ca
asishm@cs.uwindsor.ca

December 30, 2014

1 Introduction

Protein structure comparison has long been under investigation by computational biologists in hopes for finding a better but not necessarily faster alternative to the more familiar multiple sequence alignment problem. The multiple alignment problem is more challenging than pairwise alignment even for sequences, and we resort to heuristics to find as best an approximation as possible, in polynomial time. Since officially, as of 2014, there are more protein structures in the Protein Data Bank[3] than seconds in a day, there is a constant need for both speed and precision when aligning more than two proteins.

Multiple Structure Alignment (MStA) of protein structures can be categorized into four widely different approaches - progressive alignment, core optimization, graph based, and pivot based. Mustang [13], msTali [23], mulPBA [14], and CE-MC [7] use the progressive alignment approach that creates an alignment of alignments following a guide tree. While this approach does make sense, it suffers from the natural disadvantages of all progressive techniques. Methods from other approaches, [17] [29], often outperform the progressive ones, both in terms of speed and accuracy. A second way is to optimize a consensus structure, sometimes with several iterations, and report a common core of the input proteins. The idea is to find out a structurally conserved subset of residues among the proteins to gain some insight into their origin. However, such cores are mostly pseudo-structures that, although

*Research supported by an NSERC Discovery Grant to this author.

geometrically interesting, may or may not have any biological relevance. Matt [17], Multiprot [22], Mass [5], Mapsci [27], and Smolign [24] belong to this category of MStA algorithms. Ye and Godzik’s graph based POSA [29] takes a totally different path by representing a protein as a directed acyclic graph (DAG) of residues connected in the order following the backbone. POSA then creates a combined non-planar multi-dimensional DAG, taking hinge rotation into account, to come up with residue equivalences among the input proteins. While POSA gains the upper hand in terms of flexibility of protein structures, it is known to completely miss motifs on TIM barrel and helix bundles proteins [24] and incur higher cost of alignment than MATT and Smolign [24]. The pivot based approach selects one of the input molecules as ‘closest’ to all the other proteins and names it the pivot. The remaining proteins are then iteratively aligned to the pivot either in a bottom-up manner [26], or in a top-down manner [28] to come up with residue-residue correspondences that are later used to minimize some objective function and derive a score as a similarity measure. Some of the only few published algorithms in this category, are Mistral [18], [26], and [28]. Our approach is an application of the center-star method of producing a multiple sequence alignment (MSA) in an MStA scenario.

In this paper we introduce a new algorithm, **MASCOT**, which stands for **M**ultiple **A**lignment of **S**tructures using **C**enter **O**f **p**ro**T**eins, for aligning more than two proteins. Our algorithm relies on taking advantage of the linear nature of the protein polypeptide chain, while judiciously preserving the secondary structure elements (SSEs). The justification for this approach is that SSEs are latent ingredients of protein structures, serving as a well-preserved scaffold. As a result, SSEs are evolutionarily remarkably conserved while changes happen in the loops, thus modifying functionality, e.g. the substrate specificity of different serine proteases is governed by the conformation of the binding loops [9]. Further, representing protein structures by their SSEs has been successfully used on several occasions in pairwise alignment ([12], [1], [2]; [6], [15]). Our goal was to develop an algorithm that uses these sequences of SSEs and produces a multiple alignment with minimal running time and comparable accuracy. To this end, we have implemented MASCOT as a hybrid algorithm that uses a center protein, derived by minimizing sum of pairwise distance, to drive a layout that identifies a set of residues from each protein that are structurally similar. We then proceed to find an optimal correspondence among the backbone carbon atoms of these molecular structures, using inter-residue Euclidean distance threshold, and report the centerRMSD of the structures aligned in space as a measure of similarity.

2 Method

2.1 Input data set

Protein structures are stored as PDB files in Protein Data Bank [3], which contains more than 99000 structures and is regularly updated. The data can come from standard protein databases available, or from a local repository. Either way this section retrieves the correct data and makes it readily available for the algorithm to proceed. This may seem trivial at first, but for all practical purposes where thousands of macromolecules could be aligned together, one must make take care that all the requisite molecules are fetched and ready for

the next step. As an example, an input set could be 1AOR:A 7ACN 2ACT 1TTQ:B etc. As we can see, with inputs having such varied insignia, to supply the right data to the algorithm we need this first step.

2.2 Representing the proteins

Once we are assured access to the correct input set, we take each protein in turn and represent them in a way which makes further processing convenient while keeping all vital information intact.

This is a key step in the process of obtaining a useful multiple alignment due to the following factors:-

- (a) Too simple a representation could potentially miss crucial structural and functional information, whereas too complex a representation will demand innovative methods at every turn; primary sequence is the simplest possible representation and its known that they do not necessarily determine functionality. POSA on the other hand uses partial order graphs, while mulPBA uses PROFIT elements.
- (b) Difference between equal and unequal length proteins; for equal length proteins we can represent the proteins by their coordinates and apply Umeyama's method[20]. Janardan's method[28] handles unequal length proteins by representing the protein as a set of vectors using gap vectors for a gap alignment.
- (c) Choice of representation greatly affects performance;

Thus, to strike a balance between complexity and functionality, we represent the molecules as their DSSP sequences [10], which assigns each residue to one of eight possible structural motifs in a protein. Simply put, we use the linear polymer nature of the protein and stretch it out into a straight line keeping SSEs intact; which is sufficient for biological purposes.

For example a dssp representation of SEA CUCUMBER CAUDINA (1HLM) would look like this:

...HHHHGGGZZIIITTHHHHHHTTSSI...

For N input proteins we get N such dssp representations.

The main advantage of this is that we are now free to use a host of pattern matching algorithms that can compute an optimal alignment given any two such sequences - an observation which we shall exploit in the next step.

2.3 Pairwise global alignment

The processing begins by applying global alignment between every pair of dssp sequences. To do this we apply Needleman Wunsch algorithm [19] with appropriate affine gap opening penalty, and the following scoring matrix:

–	H	B	G	E	T	S	I	Z
H	1	0	1	0	0	0	1	0
B	0	1	0	1	0	0	0	0
G	1	0	1	0	0	0	1	0
E	0	1	0	1	0	0	0	0
T	0	0	0	0	1	1	0	0
S	0	0	0	0	1	1	0	0
I	1	0	1	0	0	0	1	0
Z	0	0	0	0	0	0	0	1

<i>Symbol</i>	<i>Motif</i>
H	Alpha Helix
B	Beta bridge
G	Helix 3
E	Beta strand
T	Turn
S	Bend
I	Helix 5
Z	No motif

These pairwise alignments produce a primitive picture of SSE SSE alignments. For example an alignment between dssp sequences of globins 1DM1, 1MBC, 1MBA produce the following output.

```

1DM1  ..ZZZHHHHHHHHHHHHHHHHHTHHHHHHHHHHHHHHHSGGG-...
1MBA  ..ZZZHHHHHHHHHHHHHHHHHT-HHHHHHHHHHHHHHHZGGG...

1DM1  ..ZZZHHHHHHHHHHHHHHHHHTHHHHHHHHHHHHHHH-SGGG...
1MBC  ..ZZZHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHZTHHH...

1MBC  ..ZZZHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHZTHH-H...
1MBA  ..ZZZHHHHHHHHHHHHHHHTHHHHHHHHHHHHHHHHZGGGGG...
```

We can see that the helices are properly aligned against one another. These alignments are saved in a list and referred to when needed.

2.4 Center protein

Keeping with the center star method [8] for multiple sequence alignment, an NxN symmetric matrix is created where the entries are edit distances between every aligned pair in the list mentioned above.

–	P_1	P_2	...	P_N
P_1	0	10	...	20
P_2	10	0	...	30
\vdots	\vdots
P_N	20	30	...	0

From this matrix we find the sum of pair score and the center protein using the following equation.

$$P_c = \text{protein with minimum } \sum_{i \neq j}^N \text{EditDistance}(P_i, P_j) \quad (1)$$

The protein having the minimum sum of pair score is chosen as the center protein with respect to which other proteins can be aligned. A pairwise edit distance matrix for globins 1DM1, 1MBC, 1MBA is given below:

Table 1: Pairwise Distance matrix

	1DM1	1MBC	1MBA
1DM1	0	37	4
1MBC	34	0	35
1MBA	7	35	0

The protein, at index c , having the minimum SP distance is labelled as the center protein P_c and its dssp sequence as S_c . Among the three proteins above 1DM1 is selected as the center protein since it has the lowest SP distance (equal to 41).

A lot of heuristics go on to calculate a consensus structure in place of a real protein, in hopes of finding a common core. Mustang, MultiProt, Janaradan, all attempt to procure a template structure to drive their alignment process. However, it should be noted that a consensus structure, while geometrically and perhaps computationally convenient, can often turn out to be a pseudo structure that may or may not be of any actual biological significance. Thus, our choice of going along with an actual protein to accelerate the alignment, can be deemed appropriate.

2.5 Correspondence matrix

Formally put, a correspondence matrix is an $N \times l$ matrix with respect to a center protein P_c where,

$N=|P|, P = \{P_1, P_2, \dots, P_N\}$ being the set of proteins, and

$$\max(|P_1|, |P_2|, \dots, |P_N|) \leq l \leq |P_1| + |P_2| + \dots + |P_N|$$

A correspondence matrix using the above guidelines will have the following properties:

- (a) The i^{th} row contains the ordered set of residues from protein P_i , with gaps in between.
- (b) No column will have all gaps
- (c) Gives a good idea of residue equivalences to work with, should we have to apply rigid body superposition.

In this step, all the alignments, between the center protein and every other protein, are retrieved from the list and merged one by one using the following algorithm:

Algorithm: CORRESPONDENCE MATRIX

Require: Protein DSSP sequences S_1, S_2 upto S_N **Ensure:** MSA of Sequences S_1 to S_N **for all** $i = 1 \dots N - 1$ **do** use alignment (S_c, S_i) and $\text{MSA}(S_c, S_{i-1})$ to obtain $\text{MSA}(S_c, S_i)$ following 'once a gap, always a gap' rule

A sample correspondence matrix for the globin family is given below. We can clearly see how the SSEs of all the proteins are aligned together in a column-wise fashion.

```
#####HTT-S--ZZZ--ZSSZ-Z--TTZ-Z--EEEEEEEEEE-EBSSZZSTTZEIEEEZSSS--ZEEEEEEEEEEEEEEEEETTT-TEEEEEEETT--EEEEEE--Z--TT--Z#####H-Z-H-#####T-Z|
#####HTT-S--ZZZ--SZSZ-Z--TTZ-Z--EEEEEEEEEE-EBSSZZSSZEEEEZSSS--ZEEEEEEEEEEEEEEEEETTT-TEEEEEEETT--EEEEEE--Z--TT--Z#####H-Z-H-#####T-Z
#####HTT-S--SSZSSZTSS--TTE-E--EEEEEEEEEE-EBSSZZGGGEEEEEESSS--ZEEEEEEEEEEEEEEEEGGG-TEEEEEEETTSSSEEEEESSSZ--SSSSZ#####H-Z-H-#####-Z
#####HTTSS--ZZZ--SZTSSZ--TTZ-Z--EEEEEEEEEEZEEE-ZZGGGEEEEEEETT--EEEEEEEEEEEEEEEEGGG-TEEEEEEETTSS-EEEEEE--S--STT--Z#####H-Z-H-#####-ST
#####HTTS--ZZZ--ZSSZ-ZZTTE-E--EEEEEEEEEE-EBSSZZGGGEEEEZ-ZSSSZZEEEEEEEEEEEEETTT-TEEEEEEBSSS--ZEEEEEE--Z--TT--T#####H-Z-H-#####-HH
#####HTT-S--ZZZ--SZSZ-Z--TTZ-Z--EEEEEEEEEE-EBSSZZGGGEEEEZSSS--ZEEEEEEEEEEEEEEEEGGG-TEEEEEEBSSS--ZEEEEEE--Z--TT--Z#####H-Z-H-#####-HZ
#####HTT-SS--ZZZ--ZZ-ZZ-Z--TTZ-Z--EEEEEEEEZZ-EBSSZZGGGEEEEEESSS--ZEEEEEEEEEEZEEZTTT-EEEEEESS--EEEEEE--Z--TT--Z#####H-Z-H-#####-TZ
#####HTTT-S--ZZZ--SZSZ-Z--TTZS--ZEEZZZBZZ-Z-SSZZSSZZZZZBZZSSS--ZBZZBZZZEEZZZZZSTT-ZBZZZZZSSS--ZEEZBZ--Z--SSS--ZZTTTTSZ-T-#####-SZ
TTSTTTSSSSS-SSSZSZ--ZSSZ-Z--SBZ-Z--EEZZZZZ--SSZZBTTZZBZZBZSSS--SBZZBZZEEZZZZBZZTTTSSBZZBZZ-SSS--EEZZZZ--Z--SS--ZSGGTTTZZSH-#####-ZZ
#####HTT-S--ZZSSZSTT-Z--TTZ-Z--EEEEEEEEEE-EBSSZZTTSEEEEEZSSS--ZEEEEEEEEEEEEEEEEETTT--EEEEEEBSST-TEEEEEE--ZZTTS--Z#####H-Z-H-#####-Z
#####HTT-T--ZSZ--ZZSZ-T--SSZ-TTEEEEEEEEEEE-EBSSZZTTSEEEEEZTTS--ZEEEEEEEEEEEEEEEEETTT--EEEEEEBTTSS-BEEEEEE--Z--TTS--Z#####H-Z-STT#####-Z
```

Figure 1: A correspondence matrix. [Notice there are no columns with gaps in all rows.]

At this point we have identified conserved regions across all the proteins, but not aligned them in any way. Janaradan's method reaches a similar result, creating a correspondence matrix by carefully manipulating vectors.

The result is an MSA of dssp sequences S_i for proteins P_i , $1 \leq i \leq N$. The output of this step is used to identify as many residue equivalences as possible given the raw protein structures. To actually align them in 3D space we feed this output to the next step.

2.6 Rigid body superposition

To align the proteins in space means to apply proper translation and rotation so that the distance between alpha carbon atoms of equivalent residues is below a threshold value. To do this we need two things; a set of equivalences, and a reference frame against which the rigid body superposition is to take place. We already have both these requirements fulfilled. The correspondence matrix from phase 2 gives us the residue-residue equivalences, and our chosen center protein is the reference frame. So, suppose the correspondence is as below:

Table 2: Identifying equivalences

Residue no.									
Center protein	-	-	H	H	T	I	E	-	G
Other protein	S	S	H	H	-	G	E	E	I
Residue no.	1	2	3	4		5	6	7	8

Some annotated equivalences between center and the other protein would be (1,3) (2,4) (4,5) (5,6) and (6,8).

For each protein P_i we apply Kabsch's method [11] to superpose the structures, in space, with respect to the center protein P_c . For example, our algorithm aligns globins 1DM1, 1MBC, 1MBA in this way.

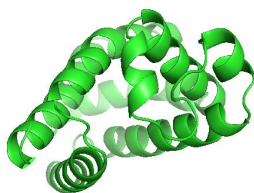


Figure 2: 1DM1

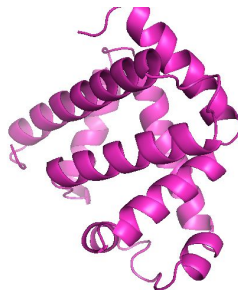


Figure 3: 1MBC

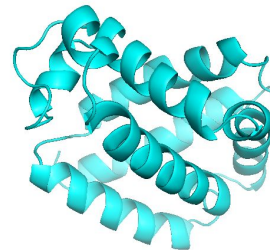


Figure 4: 1MBA

The picture above clearly shows how the structures are aligned in three dimensions.

2.7 Dynamic programming and scoring

Once the proteins are aligned in space through rigid body superposition, equivalent residues have been brought close to each other. We increase then number of equivalences between a pair of proteins by calculating the Euclidean distance between every pair of alpha carbon atoms, and considering the ones that fall within a threshold value (5Å), as equivalent pairs.

Finally we use the following formulae to derive the *centerRMSD* that represents the quality of the alignment.

$$\frac{1}{N-1} \sum_{i=1, i \neq c}^N RMSD(P_i, P_c)$$

A good alignment is one where the score is typically less than half the threshold value(2.5Å). However, difficult alignments having biological relevance can exceed this value by about 1.5Å. The centerRMSD for the alignment in Fig. 5 is 0.44Å.

2.8 Pseudocode

A pseudo-code form of our algorithm is as follows:

Algorithm: MASCOT

Require: Protein $pdbs(pdbid_1, pdbid_1, \dots, pdbid_N)$

Ensure: Multiple alignment of proteins with files created for $pdbs_{1...N}$

▷ Phase 1

- 1: Extract protein structures into $P=\{P_1 \text{ to } P_N\}$
- 2: Represent P as sequences $S=\{S_1 \text{ to } S_N\}$ consisting of DSSP SSE elements
- 3: Perform pairwise global alignment of every (S_i, S_j) using custom similarity matrix

▷ Phase 2

- 4: Create an edit distance matrix that stores the distances between every (P_i, P_j) using a custom scoring function
- 5: Choose the protein(sequence) with index c having minimum *sum of pairs* score as the center protein(sequence) $P_c(S_c)$
- 6: Create an MSA of S w.r.t S_c using center-star approach

▷ Phase 3

- 7: Treat all alignments of symbols with non-gaps as residue-residue equivalences bet. (P_i, P_j)
 - 8: Apply Kabsch's method on every (P_i, P_c) to obtain $(trans_i, rot_i)$ for every (P_i, P_c)
 - 9: Use $(trans_i, rot_i)$ from step 8 to transform and place P_i in space, with P_c being brought to origin first, to produce output pdb files
-

3 Results and Discussion

MASCOT was implemented in Python 2.7.5 using packages from Bio-python 2.0. A plethora of experiments were conducted, among which a representative set of results have been presented here. Note that T represents the time taken right from giving the input to producing the output files.

3.1 Globins

Globins are some of the most rigorously studied proteins by the MStA community. The globin family has long been known from studies of approximately 150-residue proteins such

as vertebrate myoglobins and haemoglobins. The following globins have been aligned using MASCOT:

Table 3: The table below shows the globins used in this section:

Name	PDB ids	Count	T
Set 1	1HHO:A 2DHB:A 2DHB:B 1HHO:B 1MBD 1DLW 1DLY 1ECO 1IDR:A 2LH7	10	23s
Set 2	1MBC 1MBA 1DM1 1HLM 2LHB 2FAL 1HBG 1FLP 1ECA 1ASH	10	24s
Set 3	5MBN 1ECO 2HBG 2LH3 2LHB 4HHB:B 4HHB:A	7	13s
Set 4	1ASH 1ECA 1GDJ 1HLM 1MBA 1BAB:A 1EW6:A 1H97:A 1ITH:A 1SCT:A 1DLW:A 1FLP 1HBG 1LHS 1MBC 1DM1 2LHB 2FAL 1HBG 1FLP	20	1m 38s

Set 1 is used by [18], and [24] to show how their algorithms align globins. The rmsd for this superposition is 2.765Å. Set 2, taken from [28], has been aligned with an rmsd of 2.39Å. Set 3 is [22]’s test data with rmsd 2.41Å. Set 4 is a custom assortment of 20 globins created from [16] and [27]. The purpose is to see how well they are aligned visually and with how much rmsd. As one can see, the helices and the hinges are placed within the threshold distance as much as possible, with rmsd 2.038Å.

3.2 Serpins

Serpins play an important role in the biological world. For instance, thyroxine-binding globulin is a serpine which transports hormones to various parts of the body, and Maspin is a serpine which controls gene expression of certain tumors [4]. The name Serpin stands for Serine Protease Inhibitors. The following serpins have been aligned using MASCOT:

Table 4: The table below shows the serpins used in this section:

Name	PDB ids	Count	T
Set 1	7API:A 8API:A 1HLE:A 1OVA:A 2ACH:A 9API:A 1PSI 1ATU 1KCT 1ATH:A 1ATT:A 1ANT:L 2ANT:L	13	3m 33s

The serpins in set 5 is the same one used by [22] and is said to be quite difficult owing to their large size and motif distribution. Unlike [22] we do not attempt to find a common core. Instead, we perform a global alignment over the length of the proteins. Fig. 10 shows how the beta sheets, hinges, and helices are aligned together in spite of the difficulty. Also some non-alignable parts have been correctly identified and left out. The rmsd for this alignment is 2.99Å. Fig. 11 is a low intensity PyMol rendition (LIPR) of the same alignment viewed from another angle. It uses a ribbon representation to condense the output and show most

of the aligned portions of the proteins. The pictures suggest that all these serpins share functionality and purpose, within the body. We can club all these proteins into a single family, and keep adding to it as and when such high similarities are found.

3.3 Barrels

The eight-stranded TIM-barrel is found in a lot of enzymes, but the evolutionary history of this family has been the subject of rigorous debate. The ancestry of this family is still a mystery. Aligning TIM-barrel proteins will allow us to add to this ever-expanding family. The proteins aligned in this category are as follows:

Table 5: The table below shows the barrels used in this section:

Name	PDB ids	Count	T
Set 6	1A49:A 1A49:B 1A49:C 1A49:D 1A49:E 1A49:F 1A49:G 1A49:H 1A5U:A 1A5U:B 1A5U:C 1A5U:D 1A5U:E 1A5U:F 1A5U:G 1A5U:H 1AQF:A 1AQF:B 1AQF:C 1AQF:D 1AQF:E 1AQF:F 1AQF:G 1AQF:H 1F3X:A 1F3X:B 1F3X:C 1F3X:D 1F3X:E 1F3X:F 1F3X:G 1F3X:H 1PKN 1F3W:A 1F3W:B 1F3W:C 1F3W:D 1F3W:E 1F3W:F 1F3W:G 1F3W:H 1PKM 1PKL:A 1PKL:B 1PKL:C 1PKL:D 1PKL:E 1PKL:F 1PKL:G 1PKL:H 1A3W:A 1A3W:B 1A3X:A 1A3X:B 1E0T:A 1E0T:B 1E0T:C 1E0T:D 1PKY:A 1PKY:B 1PKY:C 1PKY:D7 1E0U:A 1E0U:B 1E0U:C 1E0U:D	66	2h 25m
Set 7	1SW3:A 1SW3:B 1WYI:A 1WYI:B 2JK2:A 2JK2:B 1R2T:A 1R2T:B 1R2R:A 1R2R:B 1M5W:A 1M5W:B 1M5W:C	13	1m 22s

MASS [5] has used the 66 molecules in set 6 to show how it aligns proteins with barrels. MASCOT produces an rmsd of 3.4\AA for this alignment. Fig. 12 shows how the new algorithm can superimpose proteins having the TIM barrel supermotifs. Fig. 13 is an LIPR of the same alignment, for convenience. The result indicates these proteins have structurally highly conserved regions since all 8 helices and 8 beta sheets have been aligned. Set 7 has been taken from the gold standard manually curated SCOP database. The proteins are taken from different superfamilies but, as Fig. 14 suggests, MASCOT is still able to align the barrel motifs on top of each other, with an rmsd of 3.76\AA .

3.4 Twilight-zone proteins

Sequence alignment is still an option except when proteins have less than 30% sequence identity. The lesser the sequence similarity, the more important becomes structural comparison. Here we have taken some data sets that belong to the twilight zone.

Table 6: The table below shows the sets used in this section:

Name	PDB ids	Seq. Identity	T
Set 8	1STF:I 1MOL:A 1CEW:I	<8%	1m 55s
Set 9	1BGE:A 1BGE:B 2GMF:A 2GMF:B	<12%	5s
Set 10	1NSB 2SIM 1F8E 4DGR	<20%	19s

The above 3 sets have been chosen, after numerous trials, for their significantly low sequence similarity. The motive is to show that proteins that would never have been labeled as similar, even by the most powerful MSA techniques, can be aligned using MASCOT. This is possible because the sequential representation used here consists of SSE elements and not primary residues. Set 8, 9, and 10 represent three bands of sequence identity within the twilight zone. They have rmsd of 3.61Å, 0.1Å, and 3.15Å respectively.

3.5 Pig, Malaria, Human, and Dogfish - connected?

The 'Tree of life' has sprung many branches over millennia. Could the branches for pigs, malarial parasites, humans, and dogfish have had a common root at some point of time? The structures below have been taken from these species and an alignment is sought to gain more insight:

Table 7: The table shows the sets used in this section:

Name	PDB ids	Count	T
Set 11	1MLD:A 1MLD:B 1MLD:C 1MLD:D 1T2D:A 1I0Z:A 1I0Z:B 1LDM:A	8	49s

The crystal structure of mitochondrial malate dehydrogenase from porcine heart (1MLD) contains four identical subunits[3]. Plasmodium falciparum, the causative agent of malaria, uses the protein 1T2D to enhance NAD⁺ regeneration. Incidentally this protein is being used for new anti-malarial drugs [3]. 1I0Z, a protein from Homo sapiens, is produced by the HRAS and HRAS1 genes [3]. 1LDM represents the crystal structure of M4 apo-lactate dehydrogenase from the spiny dogfish (Squalus acanthius)[3].

Figures 18 and 19 are the same alignment viewed from different angles. MASCOT finds striking similarities among these molecules with rmsd 2.885Å, indicating that at some point of time the branches for these species might indeed have had some common ancestor.

3.6 Human, Chicken, Rabbit, Yeast, and Nematode

An ensemble group of proteins have been taken from the species mentioned above. Could molecules taken from such diverse taxa be aligned to find structural similarity?

Table 8: The table below shows the sets used in this section:

Name	PDB ids	Count	T
Set 12	1SSG:A 1SSG:B 1HTI:A 1HTI:B 1R2S:A 1R2T:A 1MO0:A 1MO0:B 7TIM 3YPI	10	47s

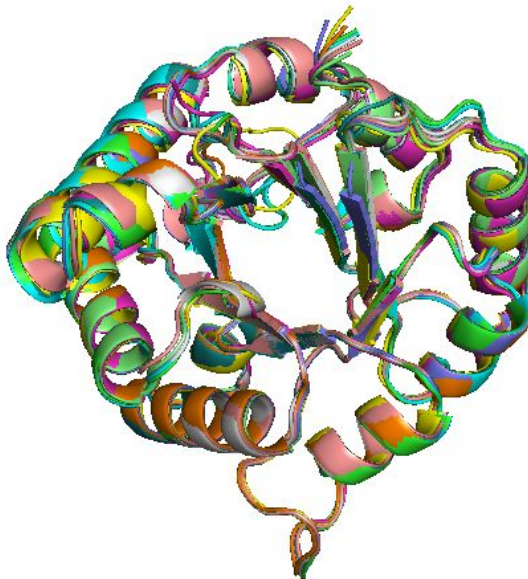


Figure 20: Set 12

Different taxa perform the same function in their own way. For example, glycolysis is the 'metabolic pathway' [21] using which glucose is broken down to form free energy. Chicken does this using the protein 1SSG [3]. Humans do the same thing using protein 1HTI[3]. We applied MASCOT to structures taken from rabbit muscle (1R2S, 1R2T), baker's yeast (7TIM, 3YPI), and nematode (1MO0), with an rmsd of 1.74\AA to confirm that these proteins are used for the same purposes.

3.7 Seafood allergy in Fish!

Rats and humans are known to have allergy towards seafood. This is generally caused due to the presence of some proteins causing havoc in the immune system. Can such propensity be exhibited among fishes?

Table 9: The table below shows the sets used in this section:

Name	PDB ids	Count	T
Set 13	1RWY:A 1RJV:A 4CPV 3PAL 1BU3 5PAL	6	5s



Figure 21: Set 12

1RWY and 1RJV are known to cause seafood allergy in common brown rats and humans. After experimenting on a host of proteins we found out some fishes too have proteins with similar structure. For example, proteins 4CPV, 3PAL, 1BU3, and 5PAL subsequently taken from common carp, pike, silver hake, and leopard shark have highly similar tertiary structures. Could this be an indication that these proteins might cause seafood allergy in these fishes? It turns out that indeed they do. A recent study by Swoboda et al [25] suggests that parvalbumins, such as the ones taken above, are major cross-reactive fish allergen. Figure 21 shows how MASCOT correctly aligns the EF hand motifs in these proteins, albeit with an rmsd of 3.82\AA .

3.8 Conclusions

This paper contributes towards the goal of comparing more than two protein structures, and finding biologically relevant similarities within them. To this end we focused on using a new approach by reducing the complexity of the three dimensional structures into meaningful SSE elements, and adopting a center-star approach to arrive at equivalences.

We have introduced MASCOT, which has been designed to overcome the major hurdles of a multiple alignment by using a sum-of-pairs heuristic that associates all proteins with the one that is 'closest' to the others among the input set.

The core of this work took the form of experiments. A representative set of results from these experiments have been presented in chapter 4. Sets 1 to 6 are standard data sets used by other published algorithms. MASCOT can efficiently align the proteins belonging to the globin, serpin, and tim barrel superfamilies. Set 7 represents data taken from a gold standard database (SCOP), which is a sort of litmus test for MStA methods. Sets 8, 9, and 10 show how MASCOT totally ignores the primary sequence and finds common motifs in spite of low sequence identity. Interesting observations have been noted through sets 11, 12, and 13. For example, set 11 shows that protein structures across these species have been conserved. So, during creation of phylogenetic trees based on structure, MASCOT can be used to process subsets of proteins as sub-problems, which later combine leading up to a tree. Set 12 and 13 are classic cases of structure-function association that suggest MASCOT results could be used to find yet-unknown biological similarities through computational methods.

4 Future Work

MASCOT can be extended to include the following functionalities in future:

1. Improve accuracy for aligning theoretical proteins.
2. Take advantage of the flexible nature of proteins and account for them when considering similarities.
3. Derive a core structure from the input proteins for use as template for protein threading.

References

- [1] Vladimir Alesker, Ruth Nussinov, and Haim J Wolfson. Detection of non-topological motifs in protein structures. *Protein engineering*, 9(12):1103–1119, 1996.
- [2] Nickolai N Alexandrov and Daniel Fischer. Analysis of topological and nontopological structural similarities in the pdb: new examples with old structures. *Proteins: Structure, Function, and Bioinformatics*, 25(3):354–365, 1996.
- [3] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
- [4] M Margarida Bernardo, Yonghong Meng, Jaron Lockett, Gregory Dyson, Alan Dombkowski, Alexander Kaplun, Xiaohua Li, Shuping Yin, Sijana Dzinic, Mary Olive, et al. Maspin reprograms the gene expression profile of prostate carcinoma cells for differentiation. *Genes & cancer*, 2(11):1009–1022, 2011.
- [5] Oranit Dror, Hadar Benyamini, Ruth Nussinov, and H Wolfson. Mass: multiple structural alignment by secondary structures. *Bioinformatics*, 19(suppl 1):i95–i104, 2003.

- [6] Helen M Grindley, Peter J Artymiuk, David W Rice, and Peter Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of molecular biology*, 229(3):707–721, 1993.
- [7] Chittibabu Guda, Sifang Lu, Eric D Scheeff, Philip E Bourne, and Ilya N Shindyalov. Ce-mc: a multiple protein structure alignment server. *Nucleic acids research*, 32(suppl 2):W100–W103, 2004.
- [8] Dan Gusfield. Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin of Mathematical Biology*, 55(1):141–154, 1993.
- [9] Lizbeth Hedstrom. Serine protease mechanism and specificity. *Chemical reviews*, 102(12):4501–4524, 2002.
- [10] W Kabsch and C Sander. Dssp: definition of secondary structure of proteins given a set of 3d coordinates. *Biopolymers*, 22:2577–2637, 1983.
- [11] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.
- [12] Ina Koch, Thomas Lengauer, and Egon Wanke. An algorithm for finding maximal common subtopologies in a set of protein structures. *Journal of computational biology*, 3(2):289–306, 1996.
- [13] Arun S Konagurthu, James C Whisstock, Peter J Stuckey, and Arthur M Lesk. Mustang: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, 64(3):559–574, 2006.
- [14] Sylvain Léonard, Agnel Praveen Joseph, Narayanaswamy Srinivasan, Jean-Christophe Gelly, and Alexandre G De Brevern. mulpba: an efficient multiple protein structure alignment method based on a structural alphabet. *Journal of Biomolecular Structure and Dynamics*, 32(4):661–668, 2014.
- [15] Guoguang Lu. Top: a new method for protein structure comparisons and similarity searches. *Journal of Applied Crystallography*, 33(1):176–183, 2000.
- [16] Dmitry Lupyan, Alejandra Leo-Macias, and Angel R Ortiz. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15):3255–3263, 2005.
- [17] Matthew Menke, Bonnie Berger, and Lenore Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology*, 4(1):e10, 2008.
- [18] Cristian Micheletti and Henri Orland. Mistral: a tool for energy-based multiple structural alignment of proteins. *Bioinformatics*, 25(20):2663–2669, 2009.
- [19] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

- [20] Satish Chandra Panigrahi and Asish Mukhopadhyay. An eigendecomposition method for protein structure alignment. In *Bioinformatics Research and Applications*, pages 24–37. Springer, 2014.
- [21] AH Romano and To Conway. Evolution of carbohydrate metabolic pathways. *Research in microbiology*, 147(6):448–455, 1996.
- [22] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. Multiprota multiple protein structural alignment algorithm. In *Algorithms in Bioinformatics*, pages 235–250. Springer, 2002.
- [23] Paul Shealy and Homayoun Valafar. Multiple structure alignment with mstali. *BMC bioinformatics*, 13(1):105, 2012.
- [24] Hong Sun, Ahmet Sacan, Hakan Ferhatosmanoglu, and Yusu Wang. Smolign: a spatial motifs-based protein multiple structural alignment method. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 9(1):249–261, 2012.
- [25] Ines Swoboda, Agnes Bugajska-Schretter, Petra Verdino, Walter Keller, Wolfgang R Sperr, Peter Valent, Rudolf Valenta, and Susanne Spitzauer. Recombinant carp parvalbumin, the major cross-reactive fish allergen: a tool for diagnosis and therapy of fish allergy. *The Journal of Immunology*, 168(9):4576–4584, 2002.
- [26] Sheng Wang and Wei-Mou Zheng. Fast multiple alignment of protein structures using conformational letter blocks. *Open Bioinformatics Journal*, 3:69–83, 2009.
- [27] Jieping Ye, Iyaylo Ilinkin, Ravi Janardan, and Adam Isom. Multiple structure alignment and consensus identification for proteins. In *Algorithms in Bioinformatics*, pages 115–125. Springer, 2006.
- [28] Jieping Ye and Ravi Janardan. Approximate multiple protein structure alignment using the sum-of-pairs distance. *Journal of Computational Biology*, 11(5):986–1000, 2004.
- [29] Yuzhen Ye and Adam Godzik. Multiple flexible structure alignment using partial order graphs. *Bioinformatics*, 21(10):2362–2369, 2005.

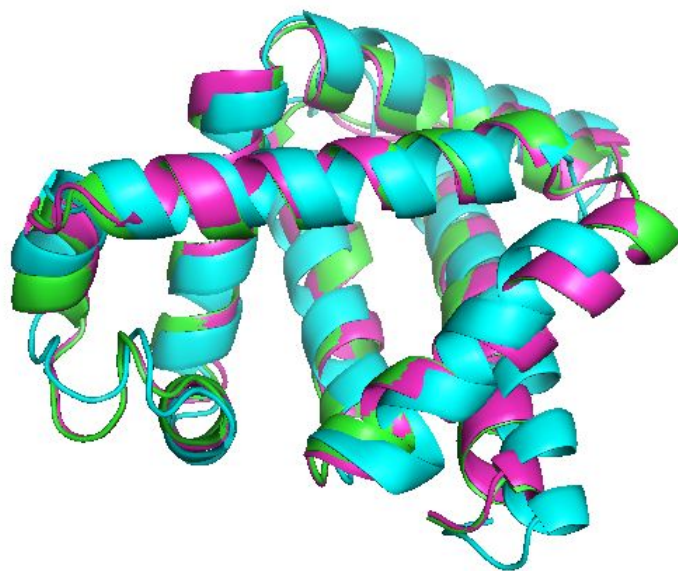


Figure 5: Alignment of 1dm1, 1mbc, and 1mba

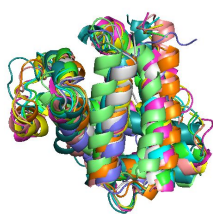


Figure 6: Set 1

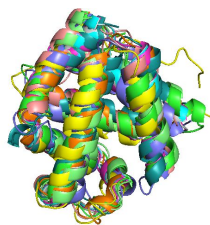


Figure 7: Set 2

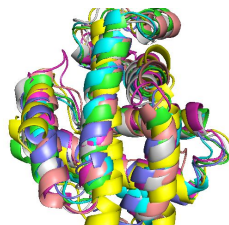


Figure 8: Set 3

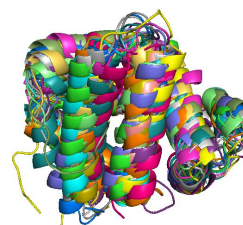


Figure 9: Set 4

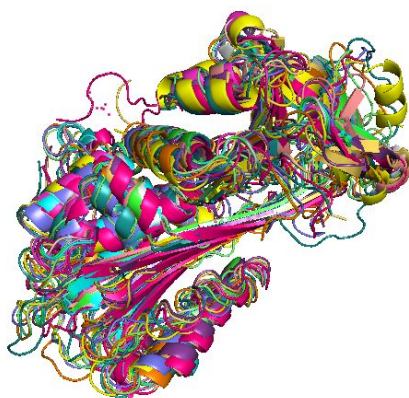


Figure 10: Set 5

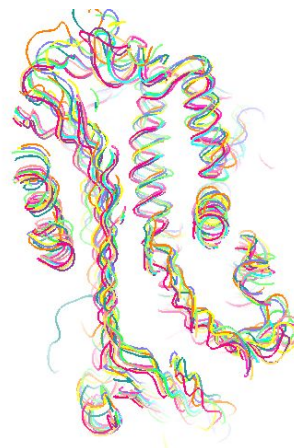


Figure 11: Set 5 LIPR

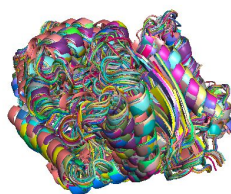


Figure 12: Set 6

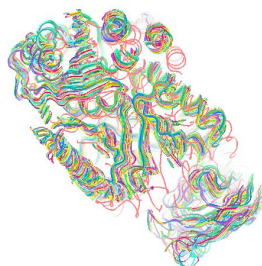


Figure 13: Set 6 LIPR

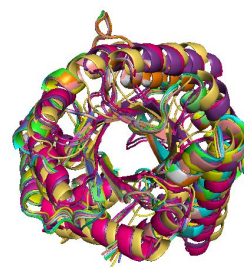


Figure 14: Set 7

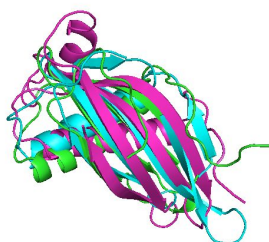


Figure 15: Set 8

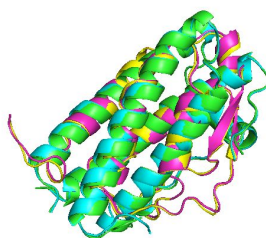


Figure 16: Set 8

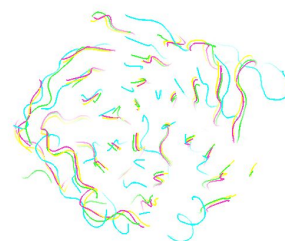


Figure 17: Set 10 LIPR

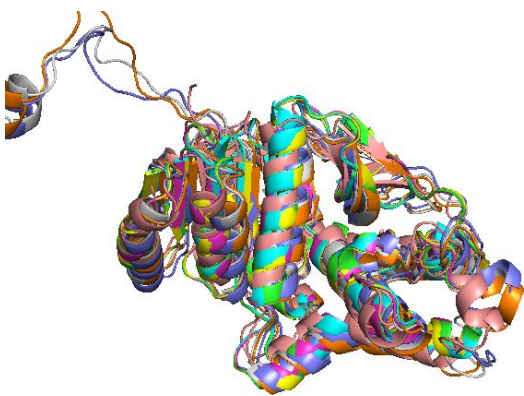


Figure 18: Set 11

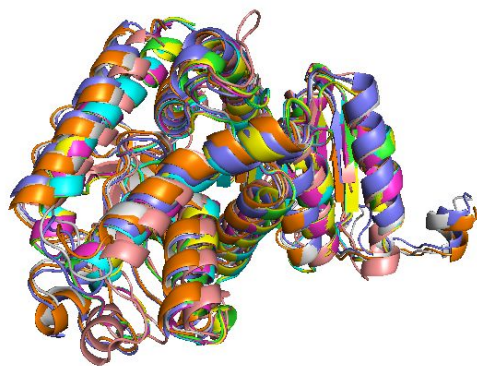


Figure 19: Set 11